社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

構造的表象からの音声生成に関する基礎的検討

齋藤　大輔† 　朝川　智† 　峯松　信明† 　広瀬　啓吉††

† 東京大学大学院新領域創成科学研究科 〒277–8561 千葉県柏市柏の葉 5–1–5
†† 東京大学大学院情報理工学系研究科 〒113–8656 東京都文京区本郷 7–3–1
E-mail: {dsk_saito,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　音声は年齢，性別，声道長や音響機器などの非言語的特徴によって不可避に歪む．筆者らはこれまでにこれらの非言語性歪みに不変な音声の構造的・抽象的表象を提案してきた．本研究では音声の構造的表象に基づく音声合成の枠組みについて提案する．提案する枠組みでは発話全体の語形（語ゲシュタルトともよばれる）を考え，それに対して身体特性，収録機器の伝送特性を与える事で初めて，聞き手が聴取する音響信号が生成される．この枠組みは，幼児の音声模倣のモデルとして解釈可能である．本報では提案する枠組みの基礎的検討として，構造的表象からの音声生成問題を制約条件下でのケプストラム空間の探索問題として定式化し，音声合成実験を行った．結果として一定の音韻性を保ち，構造抽出時の話者性ではなく，合成対象の話者性を持った音声を得ることができた．
キーワード　構造的表象，話者不変，音声模倣，言語獲得，解探索

A fundamental study of structure-to-speech conversion

Daisuke SAITO†, Satoshi ASAKAWA†, Nobuaki MINEMATSU†, and Keikichi HIROSE††

† Graduate School of Frontier Sciences, The University of Tokyo
5–1–5 Kashiwano-ha, Kashiwa-shi, Chiba 277–8561, Japan
†† Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
E-mail: {dsk_saito,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Speech acoustics vary due to differences in age, gender, vocal tract length, microphone, and so on. The authors recently have proposed a structural and abstract representation of speech, where these variations were effectively removed. In this study, a framework of speech synthesis based on this structural representation of speech is proposed. In the proposed framework, a system needs a "speech gestalt" of one utterance, properties of vocal tract length of speaker and properties of transmission of microphone. Using these information, acoustic signals to which hearers listen are generated. This framework can be regarded as a model of vocal imitation of infants. For a fundamental consideration of this framework, the authors considered this framework as a problem of searching cepstrum space for the solutions under some constraints in this report. As results of experiments, speech samples which have proper phonological characteristics were synthesized.
Key words　structural representation, speaker invariant, vocal imitation, language acquisition, searching for solutions

1. はじめに

近年の音声合成システムは，与えられたテキスト列を音響信号として出力する Text-to-Speech 変換 (TTS) が主流となっている．TTS では音韻列を音声の表象として考え，その上で漢字仮名混じり文と音韻列との対応関係，および音韻と音響信号との対応関係を統計的手法により学習する．このとき構築されるモデルは多くの場合，異音の音響モデル (triphone)，即ち音そのもののモデルである．

幼児の言語獲得のプロセスは音声模倣 (vocal imitation) と呼ばれるが，上記のように音そのものを模倣して言語を獲得する訳ではない．幼児が両親の音声の音響的実体そのものを模倣する事は声道形状の差異から不可能である．父親の「おはよう」と母親の「おはよう」を真似しても同じ本人の「おはよう」となるように，幼児は何らかの抽象化を通して音声模倣を行っていると考えられる．ここで [おはよう] という音響信号を

/おはよう/ という話者不変の音韻列に変換し，各音韻を獲得しているとの議論も可能であるが，発達心理学はこれを否定する [1]．そもそも幼児は音韻的意識が希薄であり，語から個々の音韻を抽出する能力が完成するのは小学校入学前後といわれている [2]．即ち前述した音声の抽象化と音韻による音声表象は独立であると考えられる．

幼児の音声模倣を説明しうる音声合成を考えた場合，幼児が音声模倣に際して参照している話者不変の抽象的事象を物理的，音響的に定義する必要がある．発達心理学は「幼児は単語全体の語形・音形（語ゲシュタルト [3], [4]）を獲得し，その後，個々の分節音を獲得する」と主張する [5]． 筆者らはこれまでこの「語ゲシュタルト」の音響的定義となる，話者に不変な音声の構造的かつ抽象的表象を提案してきた [6]．これは音声の音響的実体そのものは直接用いず，実体間の関係性のみをモデル化することで，非言語性歪みに対して不変性を有する音声表象である．筆者らは既にこの話者不変の音声表象を用いた音声認識システムについて検討を行ってきた [7], [8]．

本稿では，この構造的表象に基づく音声合成システムについてその枠組みを提案する．提案する枠組みは，音の実体モデルを持ちテキストを入力とする従来の音声合成システムとは大きく異なる．発話全体の語形を考え，それに対して身体特性，収録機器の伝送特性を与える事で初めて，聞き手が聴取する音響信号が生成される [9]．本枠組みは音声模倣のモデルとして解釈可能である．以下本稿ではその基礎的検討としてケプストラム空間の解探索に基づく音声合成を行い，提案する枠組みの妥当性を考察する．

2. 音声の構造的表象

2.1 非言語的特徴による音響的実体の歪み

音声の音響的実体は非言語的特徴によって不可避的に歪むが，これらは大きく乗算性歪みと線形変換性歪みに分けられる．

乗算性歪みは，スペクトルに対する乗算で表現される歪みである．ケプストラム空間では，この種の歪みは加算演算 $\mathbf{c}' = \mathbf{c} + \mathbf{b}$ として表現される．マイクロフォンの音響特性がその典型例である．また話者の声道形状差異も一部近似的に乗算性歪みであると考えられる．音声は必ず発話者を伴い，音響機器によって収録されるため，これらの歪みは不可避である．

線形変換性歪みはケプストラム空間において行列 $\mathbf{A}$ による線形変換 $\mathbf{c}' = \mathbf{A}\,\mathbf{c}$ で表現される歪みである．乗算性歪みではフォルマント周波数は変化しないが，例えば声道長差異はこれを変化させる．スペクトル表現においては，話者の声道長差異や聴取者の聴覚特性差異は周波数ウォーピングとして考えられる．周波数ウォーピングはケプストラム空間において線形変換で記述されることが示されている [10]．即ち声道長差異や聴覚特性差異は近似的に線形変換性歪みとして扱うことができる．

以上をまとめると，音声の音響的実体に不可避的に混入する非言語的特徴は，ケプストラム空間においてアフィン変換 $\mathbf{c}' = \mathbf{A}\,\mathbf{c} + \mathbf{b}$ で表現される．これらの $\mathbf{A}, \mathbf{b}$ が話者や収録環境によって多様に変化し，音声の音響的実体に様々な歪みが混入する事になる．

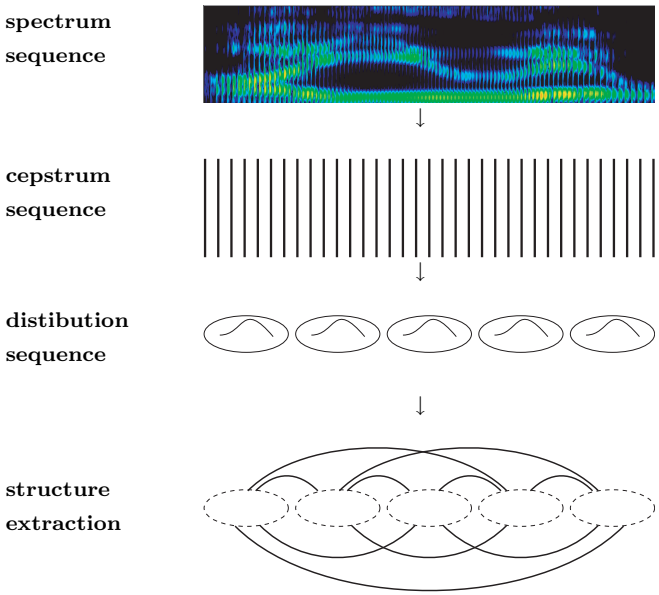


図 1　音声からの構造的表象の抽出
Fig. 1　Structure extraction from one utterance.

2.2 音声の構造的表象

ユークリッド空間において N 角形の形状は ${}_NC_2$ 個の全ての頂点間距離を規定する事で一意に定めることができる．即ち事象群に対して，全ての事象間距離を求めることでその事象群を構造的に表象することになる．しかしケプストラム空間において N 点の「**点間距離**」 によって構造を規定した場合，その構造は非言語的特徴によって不可避に歪む．なぜなら，非言語的特徴はケプストラム空間におけるアフィン変換としてモデル化され，アフィン変換は特殊な場合を除けば，構造を歪ませる変換である為である．しかしこの不可避に歪む構造は空間自体を歪ませる事で不変構造として定義することができる．

「**分布間距離**」の一つである Bhattacharyya 距離 (以下 BD と記述) を考えた場合，任意の二つの分布の確率密度関数を $p_1(x), p_2(x)$ として以下で表される．

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x) p_2(x)}\, dx \qquad (1)$$

二つの分布に対して共通のアフィン変換 $\mathbf{A}\mathbf{c} + \mathbf{b}$ を施した場合，BD は変換前後で不変となる．なおこの不変性は 1 対 1 対応のとれる非線形変換においても成立する [11]．

即ちケプストラム空間において音響事象を分布として捉え，音響事象群を「**分布間距離**」のみによって定義することで，変換不変，即ち非言語性歪みにおよそ不変な構造を求める事ができる．

2.3 一発声の構造化

一発声を一つの構造的表象で記述する場合を考える．図 1 に一発声の音声からの構造的表象の抽出の流れを示す．音声の時系列信号は，まず短時間スペクトル系列からケプストラム系列へと変換される．得られたケプストラム系列もまた時系列信号であるが，これを適当な時間区間において音響事象の分布としてとらえ，その分布の時系列へと変換する（このとき各分布に

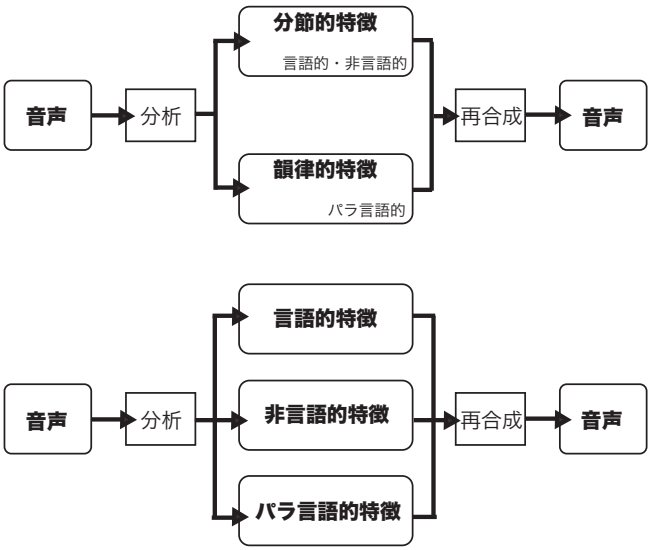


図 2　従来の分析再合成系の枠組み（上）と提案する枠組み (下)
Fig. 2　Conventional (top) and proposed (bottom) frameworks of analysis and resynthesis.

対応する時間長は分布によって異なる）．これら系列中の各分布に対して全ての組み合わせの分布間距離を求めることで一発声が構造化される．なお構造的表象はパラメータとしてケプストラムを要求する訳ではない．スペクトル包絡とケプストラムは FFT で変換でき，FFT も線形変換であるため，スペクトル表現における議論も可能である．

3.　構造的表象に基づく音声合成

3.1　非言語的要因をも分離する分析再合成系

従来の音声合成では，学習話者による数百〜数千文の音声試料を学習データとして異音（シンボル）と音との対応を学習する．そして入力としてテキスト（異音列）を与えた場合に音ストリームを出力する．この場合得られるのは学習話者の声である．

幼児の音声模倣では両親の音声を模倣しても，自らの声で言葉を返す．第 1 節で述べたように，幼児の音声模倣では両親の発声全体の語形を真似，自らの声で返していると考えられている．

本稿で提案する音声合成の枠組みは，生成対象の語形に対して，発声者の身体性（声道形状特性）を与えることで初めて音が生まれるという合成系である [9]．分析再合成系でこの考えを示すと図 2 のようになる．従来の分析再合成系では音声を分節的特徴（主にはスペクトル包絡に対応し，言語情報・非言語情報を伝搬）と韻律的特徴（主にピッチ，パワー，継続長に対応し，パラ言語情報を伝搬）に分解する．これは音声生成のソースフィルタモデルに由来する．一方提案する枠組みはこれをさらに細分化し，言語的特徴，非言語的特徴，パラ言語的特徴に分ける枠組みである．この時，言語的特徴とパラ言語的特徴を与えても音は生成されない．生成する話者は非言語的特徴の担い手であるからである．この担い手の音響特性（具体的には声道形状），更には伝送媒体のチャネル特性が与えられて初めて，

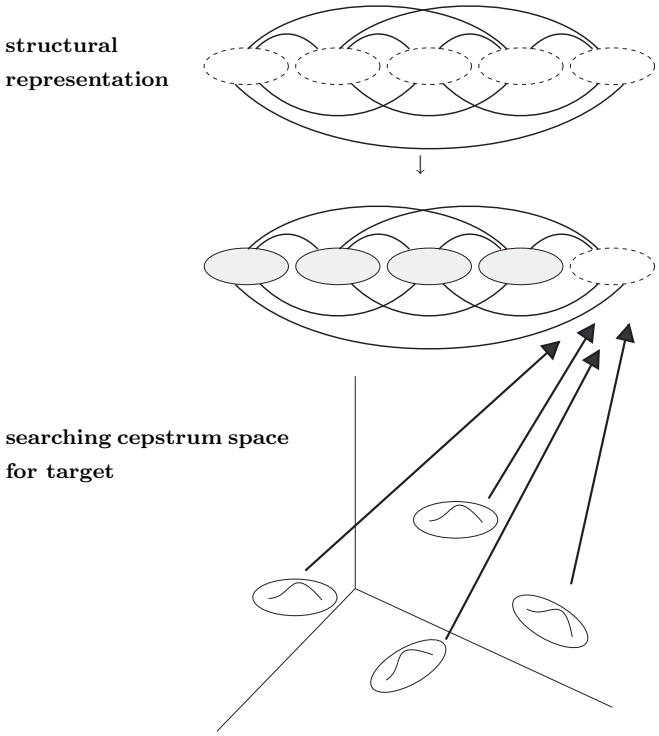


図 3　構造的表象を制約とする解探索による音声合成の枠組み
Fig. 3　A framework of speech synthesis based on searching cepstrum space under the constraint of structural representation of speech.

聞き手が聴取できる音響信号が生まれる．このことは幼児が両親の発話全体の語形を獲得し，自らの発声器官を使って言葉を発する過程をモデル化したものといえる．

3.2　ケプストラム空間の解探索

音ストリーム生成時に与える声道形状のパラメータとして調音器官の制御パラメータが考えられる．しかし調音パラメータは複雑であり，ケプストラム空間との対応関係も明確ではない [12]．そのため，今回は提案する音声合成の枠組みの基礎的検討として，ケプストラム空間における制約条件を満たす解の探索問題として定式化する．

今，構造的表象の途中までが声として出力された場面を考える（状態 $s_{t-n}, \cdots, s_{t-1}$ までが出力済み）．このとき次の出力は状態 s_t を声にする操作で得られるが，これを，$s_{t-n}, \cdots, s_{t-1}$ の音的実体 $o_{t-n}, \cdots, o_{t-1}$ からの距離制約を使ってケプストラム空間を探索して求める．このことは，構造的表象により距離関係が与えられた語形に対して，いくつかの音的実体 $o_{t-n}, \cdots, o_{t-1}$ を用いる事で，ケプストラム空間に構造を定位させていると解釈可能である．提案する解探索による音声合成の枠組みを図 3 に示す．

3.3　探索空間の制限

音響特徴量としてのケプストラムベクトルは 12〜25 次元程度の多次元ベクトルである．そのため適切な解探索のためには，探索対象となる音響空間に適切な制限を加える必要がある．今回，発声全体の平均と分散を用いて探索空間を制限することを考える．

単一のガウス分布においては平均を μ，標準偏差を σ とした時，$[\mu-3\sigma, \mu+3\sigma]$ の区間の値をとる確率は約 97%となるため，統計学ではこの区間を全空間として扱うことが多い．よって本研究では各々の次元について探索範囲をこの $\pm 3\sigma$ 区間に限定する．

上記のように探索空間を制限する事は，調音器官の運動しうる範囲の制約をケプストラム空間上で記述していることに相当する．即ち 3.2 節における構造の定位と併せれば，解探索のアプローチによっても，空間への構造定位と探索空間の制限によって，発話に身体性を与えるという提案する枠組みを示すことができることを意味している．

3.4 部分構造の歪み最小化

構造的表象は，言語的特徴と非言語的特徴を原理的に分離できる（図 2）．一方，パラ言語的特徴については一部構造的表象の中に内包される．即ちこれは発話スタイルなどのパラ言語情報によって構造が変形することを意味し，[13] では調音努力(注1)の差が構造のサイズを変化させる様子を示している．上記の探索の枠組みで音声模倣をモデル化する際，すでに出力された過去の状態がなす構造の影響を考慮しなければならない．これは幼児の音声模倣の例において，母親と幼児の調音努力が異なっていることを意味している．このときすでに出力された過去の状態が張る部分構造に母子間で差異が生じている．

構造間の差異を表す尺度について考える．構造的表象では N 個の音響事象に対して ${}_NC_2$ 個の分布間距離を求めることで構造を規定する．即ちこれらの距離の値を要素とする $N \times N$ の距離行列によって構造を表現できる．このとき距離の対称性から，距離行列の上三角部分によって構成されるベクトル（構造ベクトル）によって構造が表現できる．これらの構造ベクトルのユークリッド距離を考えることで構造間の差異を定量的に記述することができる．

上記の議論で部分構造間の差異を最小にする操作を考える．ここでは模倣対象となる構造のサイズのみを変化させて，合成対象の部分構造との差異を最小にする．このとき構造ベクトルにおいて合成対象の部分構造ベクトルの模倣対象に対する正射影を考えればよい．即ち既に音が生成された部分に関して，模倣対象の部分構造ベクトルを $\mathbf{a}$，合成対象の部分構造ベクトルを $\mathbf{b}$ として

$$\mathbf{a}' = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{|\mathbf{a}|^2}\mathbf{a} \tag{2}$$

なる操作を行う．ただし $\langle \mathbf{a}, \mathbf{b} \rangle$ はベクトルの内積を，$|a|$ はベクトルのノルムを表す．

3.5 解候補の評価

解候補の評価尺度として以下のような基準を考える．今二つの構造を表す構造ベクトルをそれぞれ $\mathbf{a}$，$\mathbf{b}$ とする．ここで二つの構造ベクトルの “**向き**” をその類似度評価に利用する．構造ベクトルの方向は，調音努力（構造サイズ）に関係なく語形を表現していると考えられ，2 つの構造ベクトルのなす角を測る事で調音努力を考慮せずに語形の類似度を評価できると考えられる．そこで類似度指標を s として

$$s = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{|\mathbf{a}||\mathbf{b}|} \quad (0 \leqq s \leqq 1) \tag{3}$$

と定義する．このとき模倣対象の語全体の構造ベクトルと，既知の状態と解候補によってできる構造ベクトルについて，この類似度 s を最大化するような候補を解として選ぶ．

(注1)：ここでは個々の音韻をよりはっきり区別しようとすることを調音努力が大きいと定義する．

表 1　音響分析条件
Table 1　Acoustic conditions.

sampling	16 bit / 16 kHz
window length	25 ms
priod	(isolated) 5 ms/ (continuous) 10 ms
parameter (isolated)	Mel cepstrum (1 to 10) [$\alpha = 0.42$]
parameter (continuous)	Mel cepstrum (1 to 12) [$\alpha = 0.55$]

3.6 特徴量空間分割

構造的表象を用いた音声認識において，構造の “過剰な不変性” のために異なる単語を同一とみなす問題が指摘されている．[8] ではこの問題の解消の為に特徴量空間を分割することで話者の違いに対してのみ適切に不変性が成立するような制約条件を導入している．また筆者らは線形変換性歪みについて，ケプストラム空間において幾何学的に強い回転性を表出することを数学的および実験的に言及している [14]．上記の特徴量分割は話者性に起因する回転に対して適切な不変性を与える効果を持つと考えられる．本研究でも余剰な解空間の排除を目的として，この手法を採用し，ケプストラムの各次元ごとに構造推定および探索を行う．

4. 合 成 実 験

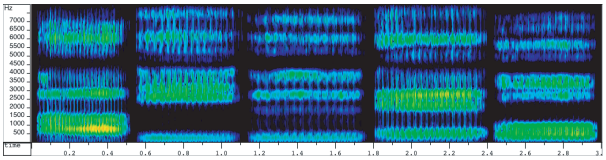
4.1 実 験 方 法

提案する枠組みによって音声合成が可能であることを確認するため，日本語 5 母音の孤立発声（/a/, /i/, /u/, /e/, /o/）および連続発声を用いて実験を行った．孤立発声に関して，成人男女各 2 名（それぞれ話者 M1，M2，F1，F2 とする）の日本語 5 母音の発声を収録した．連続発声に関しては，孤立発声とは異なる成人男女各 2 名により，各母音が一度ずつ出現する 5 母音連続発声を $5! = 120$ 通り収録した．これらの発声について表 1 に示す音響分析条件でケプストラム分析を行った．同時にそれぞれの発声のピッチ，パワー，継続長についても分析した．これらのパラメータを用いて以下に示す手順で構造抽出と解探索を行った．

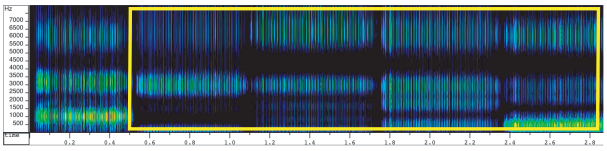
（1） 孤立発声については各母音発声を平均ベクトルと対角共分散行列で表される多次元ガウス分布として，最尤推定を行い各パラメータをモデル化する．連続発声については MAP ベースの HMM の学習アルゴリズムを利用し，ケプストラム系列から 25 状態の分布系列への変換を行う．

（2） 分布間距離を求めて構造を抽出する．

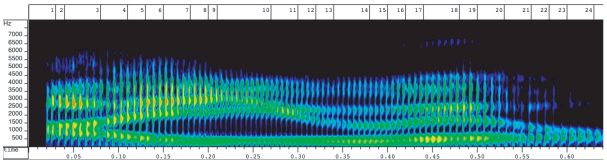
（3） 孤立発声においては 5 母音のうち n 個を既知，残りを探索対象とする．本実験では分散項は既知とし，平均ベクトルを探索範囲で変化させる．既知の母音数 n を変化させて実験を



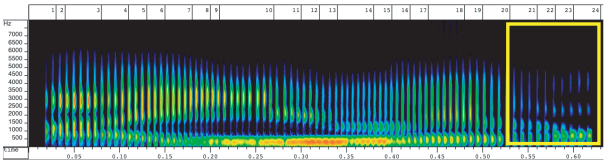
(a):原音声（孤立発声）



(b):提案手法の合成音（孤立発声）



(c):原音声（連続発声）



(d):提案手法の合成音（連続発声）

図 4　合成実験の結果．(b)(d) において枠囲いされた部分が探索による音声．

Fig. 4　Experimental results.

行う．連続発声においては 25 状態のうち 21 状態を既知とし残りの 4 状態を同様に推定する．

（4）　3.5 節の評価基準を用いて解を決定する．

（5）　得られたケプストラムの解と事前に抽出したピッチ，パワー，継続長の情報から音声を合成する．

上記の実験について，男女 1 名ずつの話者（M1，F1）を状態を生成する話者，即ち合成対象話者（音声模倣における幼児に相当）とし，構造抽出話者（音声模倣における母親に相当）を合成対象話者を含む男女 2 名としてそれぞれ実験を行った．状態生成は身体性を与える事に相当するため，構造抽出話者と合成対象話者が等しい場合は，自身の過去の発声に対して語形のみで復唱することに相当する．一方構造抽出話者と合成対象話者が異なる場合は音声模倣のモデルとして解釈可能である．

4.2　実 験 結 果

実験によって得られた合成音声の一例を図 4 に示す．図 4 には本実験に得られた音声と，比較のため合成対象話者の原音声のスペクトログラムを示している．(a) (b) は孤立発声（/a/ /i/ /u/ /e/ /o/），(c) (d) は連続発声（/aiueo/）のものである．また (a) (c) がそれぞれ原音声である．

(b) において，先頭の/a/を既知の状態として与えており，残りの 4 母音は探索によって得られたものである．スペクトルの山谷の様子を (a) と比べると，再合成の影響による平滑化はあるもののの，およそ概形を再現していることがわかる．一方 (d) においては先頭から 21 状態を既知として与えており，末尾の 4 状態は探索によって得られたものである．スペクトルの概形を (c) と比べると，末尾の/o/にあたる部分も再現されていることがわかる．ここでこの実験では (a)(c) の語形の情報は用いておらず，他者の語形情報といくつかの自身の絶対量をもとにケプストラム空間を探索し，合成すべき音声を推定していることに注意する．

5.　主観評価実験

5.1　実 験 方 法

合成実験で得られた音声について，主観評価実験を行った．今回は孤立発声の実験結果について聴取実験を行った．提案する枠組みで評価すべき項目は以下の二つであると考えられる．

表 2　聴取実験の実験条件

Table 2　Experimental conditions for the listening tests.

(a): 音韻性評価（1）の為の聴取実験条件

被験者	日本人成人男性 4 名
聴取実験サンプル	提案手法による合成音 240 サンプル
	分析再合成音による比較音声 120 サンプル
実験法	ヘッドホン聴取による書き取りテスト

(b): 話者性評価（2）の為の聴取実験条件

被験者	日本人成人男性 4 名
聴取実験サンプル	提案手法による合成音 24 サンプル
	対応する模倣対象の分析再合成音 24 サンプル
	対応する合成対象の分析再合成音 24 サンプル
実験法	上記 24 組のサンプルによる ABX 法

（1）　本手法で得られた音声について，合成すべき音韻が正しく生成されたかどうか．

（2）　本手法で得られた音声について，その話者性が構造抽出話者のものではなく身体性を与えた話者のものとして知覚されるかどうか．

（1）について評価するため，身体性を与えた話者の分析再合成音と探索によって得られた母音によって構成される 5 モーラ単語を作成し書き取りテストによる了解度評価を行った．この際「5 モーラ中に 5 母音が一度ずつ含まれる」という単語知識による影響を除くため，分析再合成音によって 5 母音の重複を許して作成した 5 モーラ単語を実験セットに加えた．さらに聴取実験に出現する単語が「日本語 5 母音による 5 モーラ単語で出現の重複は許す」と被験者に予め伝えた．

（2）について評価するために，構造抽出話者の分析再合成音および合成対象の分析再合成音を比較対象とした ABX 法により合成音声の個人性を評価した．既知母音数と構造抽出話者，合成対象話者の異なる音声サンプルを 24 通り作成し，それぞれ同一単語で ABX 法を実施した．(1), (2) それぞれの聴取実験の条件を表 2 に示す．

5.2　実 験 結 果

上記聴取実験の結果を表 3 に示す．表 3 (a) は書き取りテストの結果を示したものである．ここで正解とは 4 人の被験者のうち 3 人以上の被験者の解答と，模倣対象の単語とが一致した

表 3　聴取実験の結果
Table 3　Results of the listening tests.

(a):書き取りテストの結果

	総単語数	正答率（5 モーラが全て一致）	正答率（1 モーラの誤りを許容）
分析再合成音	120	0.67	0.83
既知母音数 1	40	0.15	0.38
既知母音数 2	80	0.20	0.44
既知母音数 3	80	0.43	0.81
既知母音数 4	40	0.48	0.95

(b):話者性評価の結果

	全員一致	3 人以上一致
全体	18/24	24/24
既知母音数 1	5/6	6/6
既知母音数 2	4/6	6/6
既知母音数 3	4/6	6/6
既知母音数 4	5/6	6/6

場合を表している．分析再合成音の結果は単語知識の解消の為に用いた，重複出現を許す 5 モーラ単語の結果である．これにより既知母音数を 1 として合成した場合でも 15 %ほどが単語完全一致で再現可能である事がわかる．さらに 1 モーラの誤りを許容して正解に加えた場合，大幅に正答率が上昇し，既知母音数が 1 の場合でも約 4 割の単語が正しく合成できた事になる．

表 3 (b) は話者性評価実験の結果を示したものである．表 3 (b) は被験者 4 名のうち 3 人以上，または全員一致で合成対象話者に近いと評価した数を示している．これにより既知母音の数に依存せず，模倣対象の話者性を除去し，合成対象の話者性を再現できている事がわかる．

6. 考　　察

合成実験，および主観評価実験の結果について考察する．表 3 (a) より，合成された単語の了解度としては既知母音数が 4 のとき 48 %と分析再合成の値 (67 %) を下回った．おもに了解度が低くなる原因として以下が考えられる．

（1） 解探索における解像度の問題

（2） 解候補の評価における局所解の影響

（1）は各次元の全探索を行う際の量子化の解像度の問題である．この解像度を高く設定する事で正しい解が得られると考えられるが，計算時間とのトレードオフとなる為，最適な値を検討する必要がある．（2）は解候補の評価において本来異なる音に対して語形が類似していると判断してしまうことを意味している．式 (3) によって得られる，構造ベクトルの “向き” の類似度による評価は，調音努力の影響を考慮しない評価尺度である．そのためある程度の発話スタイルのばらつきを抑える事ができる反面，構造をケプストラム空間に定位させる絶対量が少ない場合，調音努力が大きく異なる単語を等しいと判断する可能性を持っている．そのため構造歪みの最小化基準も併せて考慮した評価基準を検討する必要がある．また図 4 のスペクトルのように，今回は再合成において分散項を考慮していないため全体的にブザー音のような音声が出力されてしまっている．このことも了解度を下げた要因であると考えられる．

なお主観評価実験では/i/と/u/を取り違えた間違いが目立った．また音韻交代のような現象も多く見られた．

表 3 (b) により，話者性の影響がおよそ取り除かれていることがわかる．この結果より特徴量空間の分割が有効に機能していることが示唆される．一方，今回はケプストラム空間を 1 次元ずつの部分空間に分割したが，ケプストラムの各次元の相関を考慮し，最適な特徴量空間の分割を行う事で，話者性の影響を除去したうえで合成音声の品質を向上させる事ができると考えられる．

なお今回連続音声に関しては主観評価実験を行わなかったが，予備的な聴取ではおよそ正しい音韻性の音が合成されていた．

7. おわりに

本稿では，非言語性歪みにおよそ不変な特徴量である音声の構造的表象に基づく音声合成の枠組みを提案した．音声の構造的表象は話者性を含まない物理表象であるため，音声合成に際しては明示的に身体性を与える必要がある．提案する枠組みにおける音声合成を検証するため，構造的表象を制約条件とし，既に音化された事象をもとに次の音響事象を推定する形で問題を定式化した．実際にケプストラム空間の解探索によって音声合成を行い，主観評価実験によって一定の音韻性を保ち，構造抽出話者ではなく合成対象話者の話者性を持った音声が合成されたことを確認した．

提案手法は，幼児の音声模倣のモデルとして捉えることができる．合成実験において，異なる話者の構造を用いた音声合成においても，構造抽出した話者の話者性の影響を受けずに合成音が生成されており，提案する枠組みの妥当性を示唆している．

今後は分散項を考慮して再合成におけるスペクトルの連続性を解決する事や，声道形状特性を直接的に入力パラメータとする点が課題としてあげられる．さらに 6. で述べた探索条件や評価基準，最適な次元分割数などの再検討も行っていく予定である．また連続音声に対する主観評価実験や，より大規模な聴取実験によって提案する枠組みの妥当性を示していく予定である．

文　　献

[1] 内田伸子編, “発達心理学キーワード”, 有斐閣双書, 2006.
[2] 天野, “子どものかな文字の習得過程”, 秋山書店, 1986.
[3] 早川, 月刊言語,35,9,pp.62–67,2006.
[4] N. S. トルベツコイ, “音韻論の原理”, 岩波書店, 1958.
[5] 加藤, コミュニケーション障害学, 20, 2, pp.84–85, 2003.
[6] N. Minematsu et al., Proc. SRIV’2006, pp.47–52, 2006.
[7] 朝川他, 信学技報, SP2006-105, pp.119–124, 2006.
[8] 朝川他, 音講論（秋）, 3-Q-10, 2007.
[9] 峯松他, 信学技報, SP2007-30, pp.37–42, 2007.
[10] M. Pitz, H. Ney, IEEE Trans. Speech and Audio Processing, Vol.13, pp.930–944, 2005.
[11] 峯松他, 音講論（春）, 1-P-12, 2007.
[12] 錦戸, 党, 音講論（春）, 1-Q-28, 2007.
[13] N. Minematsu et al., ICASSP2006, Vol.1, pp.261–264, 2006.
[14] 齋藤他, 音講論（秋）, 1-P-2, 2007.